Efficient coding of natural sounds

Michael S. Lewicki

Computer Science Department and Center for the Neural Basis of Cognition, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213, USA
Correspondence should be addressed to M.S.L. (lewicki@cnbc.cmu.edu)

Published online: 18 March 2002, DOI: 10.1038/nn831

The auditory system encodes sound by decomposing the amplitude signal arriving at the ear into multiple frequency bands whose center frequencies and bandwidths are approximately exponential functions of the distance from the stapes. This organization is thought to result from the adaptation of cochlear mechanisms to the animal's auditory environment. Here we report that several basic auditory nerve fiber tuning properties can be accounted for by adapting a population of filter shapes to encode natural sounds efficiently. The form of the code depends on sound class, resembling a Fourier transformation when optimized for animal vocalizations and a wavelet transformation when optimized for non-biological environmental sounds. Only for the combined set does the optimal code follow scaling characteristics of physiological data. These results suggest that auditory nerve fibers encode a broad set of natural sounds in a manner consistent with information theoretic principles.

© 2002 Nature Publishing Group http://neurosci.nature.com

Much is known about how the brain encodes sensory information, but the question of why it has evolved to use particular coding strategies has long been debated[1]. In the auditory system, cochlear nerve fibers are sharply tuned to specific frequencies and can be characterized as performing a short-term spectral analysis of acoustic signals[2]. To a first approximation, the frequency and phase responses of auditory nerve fibers can be modeled as a bank of linear filters that integrate auditory information over a timescale that varies with frequency[3–5]. Although these filtering properties resemble Fourier and wavelet transforms, this observation alone is not an adequate explanation for the auditory code. It is not clear whether these transforms, which are derived largely from mathematical considerations, are appropriate for processing the sensory stimuli experienced by an organism. A tonal decomposition might seem like a natural choice for harmonic sounds such as vocalizations, but the natural environment is rich with sounds that are not harmonic. If these have equal behavioral significance, one would expect auditory systems to be adapted for processing a broad class of sounds. In this case the optimal code is less obvious.

Can auditory sensory codes be explained by theoretical principles? One view, efficient coding theory, holds that the goal of sensory coding is to encode the maximal amount of information about the stimulus by using a set of statistically independent features[1,6–9]. Auditory nerves encode naturalistic stimuli more efficiently than white noise[10], but it is not known whether the properties of the code itself can be predicted from the statistics of the environment. Testing theoretical predictions not only offers insight into the organization of the auditory neural code, but also reveals how the codes of different organisms might be adapted for different auditory environments.

Efficient coding has successfully explained the properties of receptive fields in primary visual cortex by deriving efficient visual codes from the statistics of natural images[11–14]. To test this theory in the auditory system, we used independent component analysis, to derive efficient codes for different classes of natural sounds, including animal vocalizations, environmental sounds and human speech. This predicted a theoretically optimal code and provided an explanation for both the form of the filtering properties of cochlear nerves and their organization as a population. Previous explanations based on average power spectra, which do not take temporal regularities into account, do not accurately predict the population characteristics of cochlear nerve fiber responses. Here, the sound class that yielded a code most similar to physiological observations was a mixture of environmental sounds and animal vocalizations. Identical results were obtained with human speech, suggesting that its acoustic features make efficient use of the coding capacity of the auditory system.

Results

Auditory coding model

An auditory code based on the information theoretic principle of efficient coding can be derived by assuming the signal $x(t)$ in a time window of length N is encoded in a set of M responses $a_1(t),..., a_M(t)$. The goal of efficient coding is to derive a set of filters $h_1(t),..., h_M(t)$ that minimize the statistical dependence of the responses[6–9]. The response of a particular filter i is

$$a_i(t) = \sum_{\tau=0}^{N-1} x(\tau) h_i(t - \tau) \qquad (1)$$

Any statistical dependence among the filter outputs implies that two different channels are transmitting the same information. Redundancy can be an advantage in the presence of noise (arising from the auditory signal itself or from imprecise coding) because it adds robustness. However, if the signal-to-noise ratio is high, as in the current case, redundancy among the channels means that the bandwidth is not optimally utilized. In theory, an ideal code transforms the input signals so that the outputs are statistically independent, removing all redundancy. Current methods can only approximate this ideal within the limits of the

© 2002 Nature Publishing Group http://neurosci.nature.com

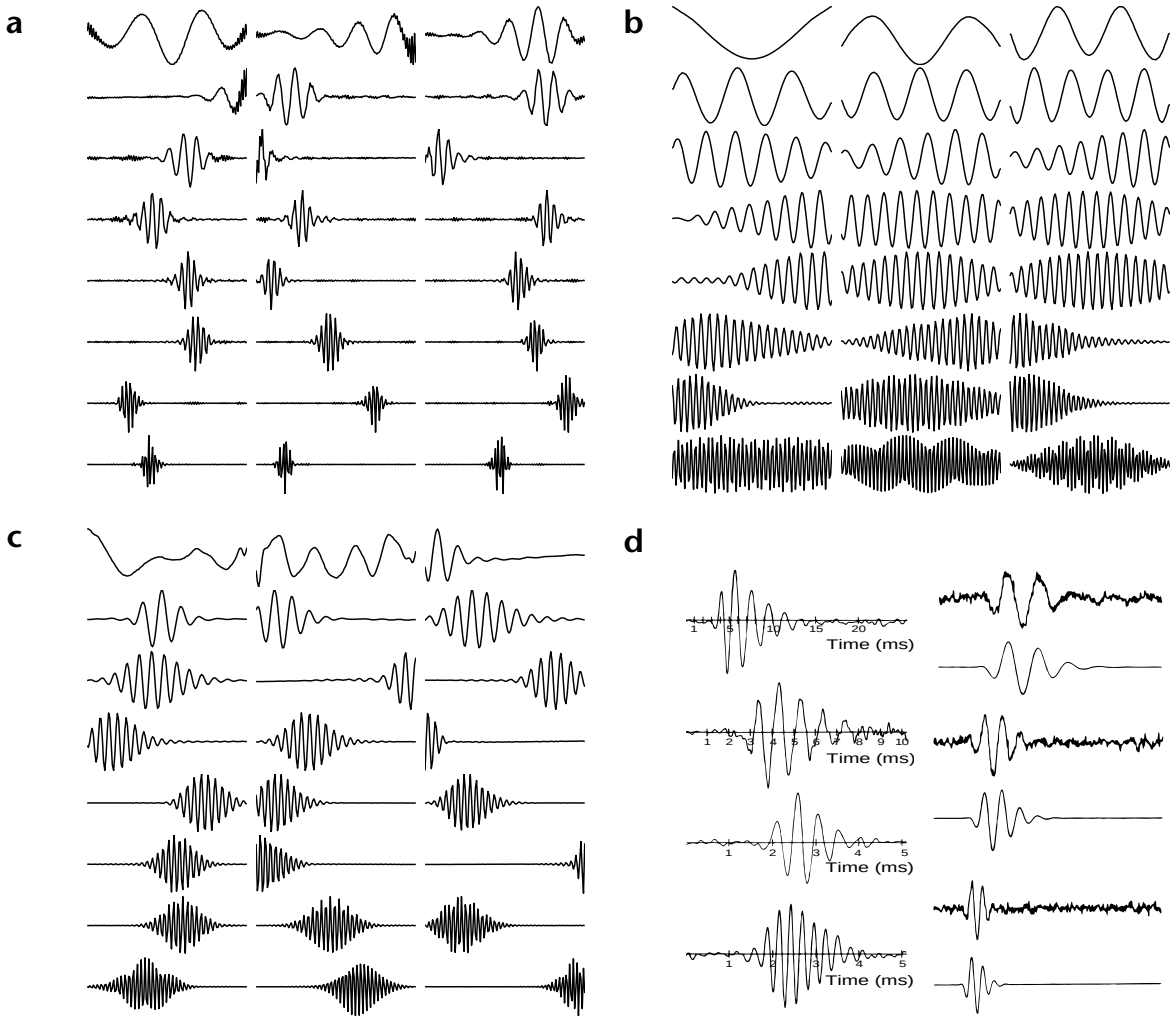


Fig. 1. Auditory filters derived from efficient coding of different natural sounds classes. Individual waveforms show the entire filter in the time domain (~8 ms). The set of filter shapes is optimized to form an efficient code by maximizing the statistical independence of their response over the sound ensemble. Each plot shows a representative subset of the total population of 128 filters, displayed in increasing order of peak resonance frequency. (**a**) Efficient coding of non-harmonic environmental sounds yields a set of filters that resemble a wavelet representation. The majority have a dominant resonance frequency (see **Fig. 2**) and an amplitude envelope that is localized in time. (**b**) Efficient coding of animal vocalizations results in filters that resemble a Fourier representation. All filters are sinusoidal and the majority extend over the entire length of the analysis window. The moiré-like patterns visible at the highest frequencies arise from the cyclic alignment of the underlying filter resonance frequency and the sampling frequency. (**c**) Efficient coding of speech, which contains both harmonic and non-harmonic sounds (that is, vowels and consonants), yields a representation intermediate between those in (**a**) and (**b**). (**d**) Comparison to cochlear filter shapes measured experimentally at the auditory nerve. The physiological filters are redrawn from original figures[4,5]. Left (from ref. 4), filters with peak resonance frequencies of 0.53, 1.0, 2.1 and 4.7 kHz from top to bottom (note different time scales). Right (from ref. 5), measured filters (upper) and modeled functions (lower). Each waveform is 20 ms in duration; the peak resonance frequencies are 364, 642 and 999 Hz. Like the filter shapes predicted by efficient coding, the auditory nerve filters integrate the auditory signal over a time period that depends on frequency.

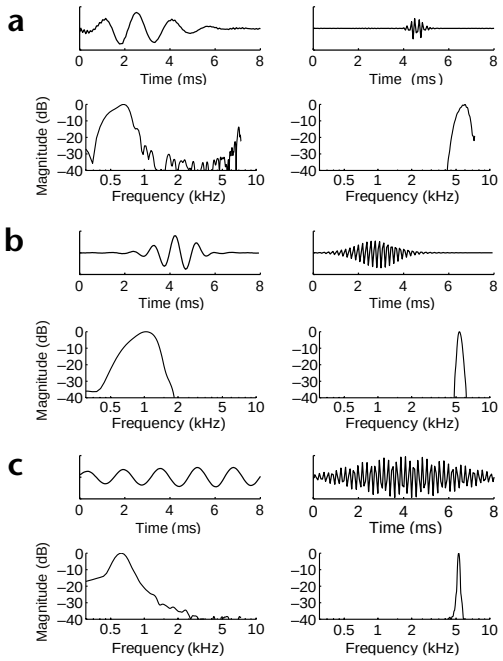
assumed model. Methods for deriving efficient codes for models of the form in equation (1) fall under the rubric of either sparse coding[11] or independent component analysis (ICA)[15,16], and are aimed at finding the features (or basis functions) that model the statistical distribution of the pattern ensemble[14].

Predicting codes for natural sounds

The notion of an efficient code cannot be separated from the ensemble of signals that are being encoded[6,17]. To make predictions for sensory codes, it is necessary to make conjectures about what class of stimuli the sensory system has evolved to process. This could range from a broad class of signals in the natural environment to only those crucial for reproduction and survival. Many auditory systems, such as those of barn owns and bats, have highly-specialized adaptations. The goal here, however, was to make predictions about less specialized, 'general' auditory systems. We therefore chose to analyze three classes of sounds as representatives of a natural auditory environment—environmental sounds, animal vocalizations and human speech—with each class containing a broad array of different sounds, animals or speakers (see Methods). Environmental sounds, such as rustling brush, crunching leaves and snapping twigs, call for rapid and accurate auditory localization. These sounds are typically broadband, non-harmonic and of short duration. Animal

© 2002 Nature Publishing Group http://neurosci.nature.com

Fig. 2. Filter power spectra. Time domain plot (upper) and corresponding power spectrum (lower) of representative filters derived from environmental sounds (**a**), human speech (**b**) and animal vocalizations (**c**).



vocalizations, which have likely evolved to stand out from background sounds, are typically harmonic, of relatively narrow bandwidth and of longer duration. Speech, the third distinct class of sounds, is highly relevant to the human auditory system and shares properties of both animal vocalizations and environmental sounds by virtue of having both harmonic vowels and non-harmonic consonants.

Efficient codes for each sound ensemble were derived using a generalized independent-components algorithm (see Methods). The analysis window was limited to 128 samples. At a sampling rate of 14.7 kHz for vocalizations and environmental sounds and 16 kHz for speech (see Methods), this corresponds to a window width of approximately 8 ms. This width was chosen for computational efficiency and because it covers the relevant time scale for a broad range of auditory nerve fibers. This window width also captures most of the short-range temporal correlations in the sound ensembles as revealed by the auto correlation functions (data not shown), although longer-range and higher-order statistical relationships clearly exist. Each filter is defined by 128 points over the analysis window, and the algorithm places no constraints on filter shape. That is, each of the 128 points that define the filter is a free parameter, so the filters could take on any spectral or temporal pattern. The shapes that emerge are determined by the statistical structure of the ensemble. For environmental sounds, even though the ensemble consisted of non-harmonic sounds like rustling brush and cracking twigs, the majority of the filters have a single peak resonance frequency and an amplitude envelope that is localized in time (**Fig. 1a**). Similar filter shapes can appear at multiple temporal positions, because the set of filters is optimized to encode the waveform over the whole analysis window.

The properties of the filter population change with the statistical structure of the sound ensemble. The ICA filter shapes for animal vocalizations are essentially Fourier representations, with most filters having sinusoidal oscillations and little amplitude modulation (**Fig. 1b**). This type of representation was expected, as the sounds in the ensemble of animal vocalizations were largely harmonic. The filters are not localized in time because the statistical regularities of animal vocalizations occur on a much larger timescale than those of environmental sounds. A much larger analysis window would be required to show the temporal extent of this regularity. Efficient coding of speech also yields sinusoidal filters, but with amplitude envelopes that lie between those for environmental sounds and animal vocalizations (**Fig. 1c**). Efficient coding of sounds can produce filters that are localized in time and frequency[18,22]. The trade-off between time and frequency is clearly visible in the progression from environmental sounds to vocalizations, reflecting the correlation time of the different sound ensembles (**Fig. 2**). This pattern was not obvious from the average autocorrelation of the three sound ensembles (unpublished data).

The filters derived from efficient coding are qualitatively similar to filters that model the response properties of auditory nerve fibers (**Fig. 1d**). The auditory nerve filters were estimated using reverse correlation ('revcor' filters), which provides an estimate of linear filters that determine both the temporal and spectral properties of the auditory nerve response[3–5]. The revcor filter shapes resemble those derived theoretically for environmental sounds and human speech, and also show the dependence of amplitude envelope width on frequency. Similar filter shapes also account for the spectral analysis properties of the human auditory system[19].

These results are consistent with the notion that the sensory code of the auditory system is efficient for a mix of non-harmonic, broadband sounds and harmonic vocalizations. But it is possible that the intermediate temporal envelopes observed for speech arose because of a particular acoustic property of speech, not because it contains both harmonic and non-harmonic sounds. To test this explicitly, the same analysis was performed on a combined sound ensemble consisting of environmental sounds and animal vocalizations. The relative proportion of each ensemble could be chosen, and a 2:1 ratio of envi-

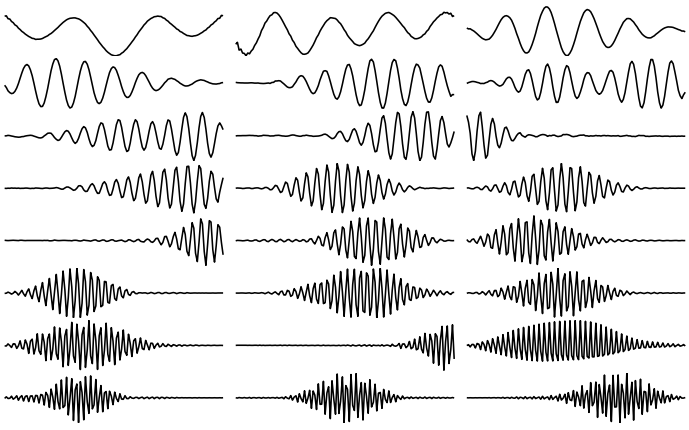


Fig. 3. Efficient coding of a combined sound ensemble. Efficient coding of a sound ensemble consisting of environmental sounds and vocalizations in a 2:1 proportion yields filters similar to those for speech, with temporal envelopes that are intermediate between those for environmental sounds and animal vocalizations.

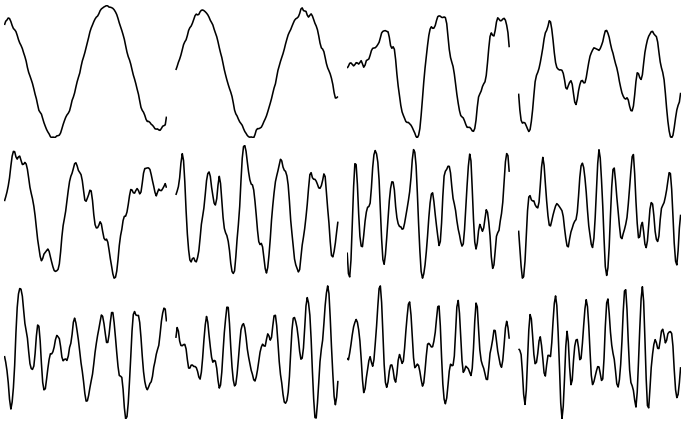


Fig. 4. Principal components of natural sounds. A representative subset of filters derived from PCA of environmental sounds are plotted in decreasing order of captured variance. PCA, which can only model second-order correlations, does not yield filters that are localized in time, and only the largest are sinusoidal.

© 2002 Nature Publishing Group http://neurosci.nature.com

ronmental sounds to animal vocalizations yielded filters similar to those for speech (**Fig. 3**).

Principal-components analysis (PCA or the Karhunen-Loéve transform) has long been used in efficient coding of speech signals[20]. To contrast PCA-derived with ICA-derived predictions, a set of filters for the same ensemble of environmental sounds was derived using PCA (**Fig. 4**). The largest principal components are sinusoidal, but most are not localized in either time or frequency and bear little resemblance to auditory filters. A Fourier-like code would be expected for sufficiently large data ensembles, because of the assumption of stationarity (that the statistical structure of the sound ensemble is not dependent on the temporal position of the analysis window). PCA selects filter shapes that decorrelate the outputs, and embodies an implicit assumption that the outputs follow a Gaussian distribution. For the datasets used here, however, the outputs are highly non-Gaussian, and decorrelation, which results in a less efficient code, is not sufficient to explain the auditory filter properties. Furthermore, the PCA filters are restricted to be mutually orthogonal, which greatly restricts the class of filters that can be used to model structure in the sound ensemble. This restriction is not imposed by ICA.

To check for a bias in the efficient coding algorithm giving wavelet-like filters, the algorithm was run on a data ensemble in which the samples were drawn independently from a sparse distribution ($p(x) \propto \exp(-|x|^{0.5})$). As expected for a data set that contains no temporal structure, the resulting filters were maximally localized in time, with each representing a different temporal position (**Fig. 5a**). If the algorithm is run on a single speaker from the speech dataset, the filters are not localized in time and adapt to encode particular harmonics of the speaker's voice (**Fig. 5b**).

Analysis and characterization of the derived codes

How the filter populations encode the three sound classes can be characterized using time–frequency analysis, or the distribution of the filters in terms of their temporal envelopes and spectral power. For example, a Fourier transform represents a signal by a linear superposition of sinusoids. Thus, the filters are localized in frequency but not in time (**Fig. 6a**). A wavelet transform, by contrast, is composed of filters that are localized in both time and frequency (**Fig. 6b**). Because the two codes cover the time–frequency space with the same number of functions, a wavelet representation sacrifices frequency resolution while improving time resolution. The tiling of time–frequency space (that is, the partitioning of time–frequency space by the filter set) plays a central role in the design of wavelet transforms[21]. Because the individual filters derived from efficient coding are localized in their amplitude envelope and their spectral power, it is possible to plot how each population covers time–frequency space. For coding purposes, the best tiling depends on the statistical structure of the signals.

Time–frequency analysis (see Methods) of the derived filters shows that the majority of filters are localized (**Fig. 2**). The time–frequency distribution for the environmental sounds code is similar to a wavelet representation, with bandwidth increasing and temporal width decreasing as function of frequency (**Fig. 6c**). One difference, however, is that instead of discrete increases in bandwidth and decreases in temporal width, as is common for many types of wavelets, bandwidth and temporal width change gradually with frequency. For animal vocalizations, the efficient code most closely resembles a Fourier representation, and the bandwidths are much narrower than the codes for the other datasets. The time–frequency distribution for the efficient code of speech falls between the two others in both temporal extent and filter bandwidth. Deriving an optimal code provides a solution to the choice of how to tile time–frequency space, but it is more general because the filters are not restricted to be localized in time–frequency space. Harmonic structure is one example of nonlocal time–frequency structure, and is obtained for an efficient code of a single speaker[22] (**Fig. 5b**). That the majority of the filters are localized reflects the statistical structure of the signals.

The differences between the time–frequency tiling for the three sound classes can be summarized by plotting characteristics

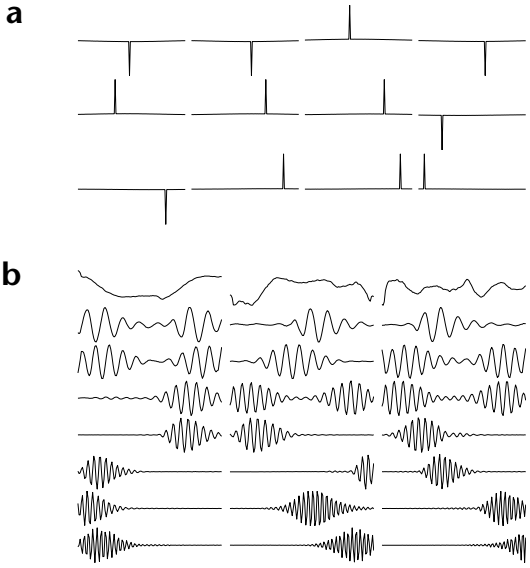


Fig. 5. Control analyses. Sinusoidal filters that are localized in frequency are not inherently preferred by the algorithm. (**a**) Applying the same algorithm to sparse noise, where there is no temporal structure, results in an impulse representation where the filters are maximally localized in time. (**b**) Applying the algorithm to speech from a single speaker results in nonlocalized filters that are adapted to the harmonics of the speaker's voice.

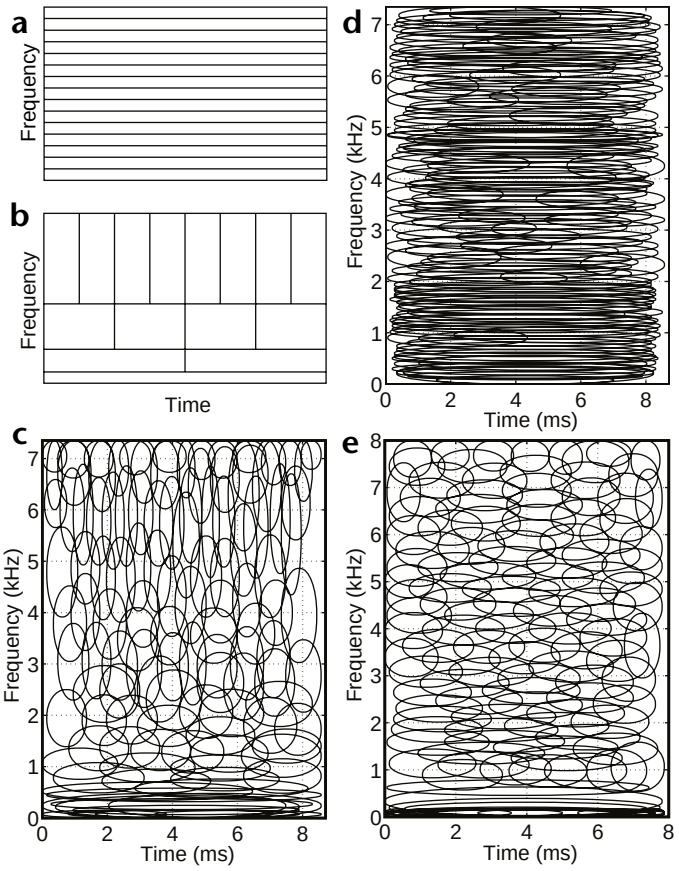


Fig. 6. Time–frequency analysis. (**a**) The filters in a Fourier transform are localized in frequency but not in time. (**b**) Wavelet filters are localized in both time and frequency. (**c**–**e**) The statistical structure of the signals determines how the filter shapes derived from efficient coding of the different data ensembles are distributed in time–frequency space. Each ellipse is a schematic of the extent of a single filter in time–frequency space. (**c**) Environmental sounds. (**d**) Animal vocalizations. (**e**) Speech.

of the filters in time–frequency space as a function of center frequency (**Fig. 7**). For comparison to auditory nerve filtering properties derived physiologically and psychophysically, we analyze bandwidth, filter sharpness (center frequency divided by bandwidth, or *Q*) and the temporal envelope[23]. Bandwidth is measured 10 dB down from the spectral peak. Filters that did not have a full 10 dB drop on both sides of the spectral peak (the lowest and highest frequencies) were omitted from the plots to avoid artifacts resulting from the limited size of the analysis window. The filters optimized for environmental sounds show the steepest increase in bandwidth as a function of frequency, similar to a wavelet representation (**Fig. 7a**). By contrast, the filters derived for vocalizations have bandwidth that is nearly constant across frequency, as in a Fourier representation. The curve for speech lies intermediate between the other two. The corresponding curves for the temporal envelope necessarily show the same pattern because of the time/frequency trade-off (**Fig. 7b**). The filters for all three sounds classes show an increase in sharpness with center frequency (**Fig. 7c**). All curves approximately follow a power law.

Deriving an efficient code for the combined set of environmental sounds and animal vocalizations (**Fig. 3**) yields similar bandwidth and sharpness curves. The curves for these filters can be shifted from one extreme to the other by changing the relative proportion of the two types of sounds in the dataset (unpublished data). The curves most consistent with physiological measurements[24,25] are those for the speech data set and the combined sound ensembles (**Fig. 7d**).

One argument for explaining the increase in filter bandwidth with center frequency is based on the observation that the average power spectrum for speech, music and some natural sounds is approximately 1/*f* (refs. 26,27). If frequency bandwidths are chosen so that each band has equal average power, the bandwidth must increase linearly with frequency[28]. The functions of equal power bandwidth versus center frequency derived from the power spectra of the natural sound ensembles do not agree as closely with physiological and psychophysical observations, and differ from those derived from efficient coding (**Fig. 8**). The vertical position of the curves in Fig. 8b is arbitrary and was adjusted to best match the range in Fig. 7a. The position is influenced by the number of assumed frequency channels: doubling the number halves the bandwidth of each channel. Auditory nerve fibers have highly overlapping

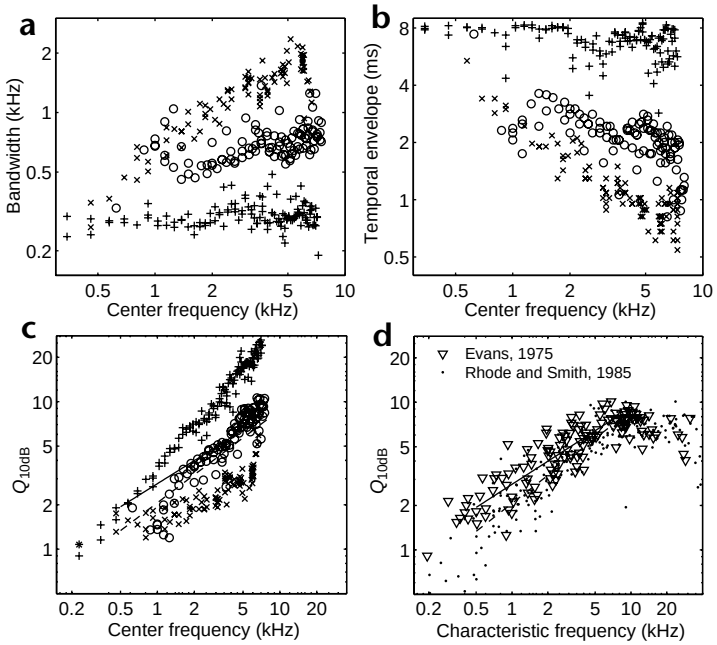


Fig. 7. Comparison of filter population characteristics to physiological data. (**a**–**c**) Characteristics of the derived filters as a function of center frequency for environmental sounds (×), speech (O), and vocalizations (+). (**a**) Filter bandwidth. (**b**) Filter temporal envelope width. (**c**) Filter sharpness or center frequency divided by bandwidth (Q_{10dB}). For comparison, the linear regression lines from the physiological data (**d**) are superimposed. The curves for a combined ensemble of environmental sounds and animal vocalizations can be varied smoothly from one extreme to the other by changing the relative proportion of the two sound classes (data not shown). (**d**) Q_{10dB} measured from cat auditory nerve fibers. Lines show linear regressions of each dataset in the range 0.5–8 kHz. Data are replotted from ref. 24 (5 and solid line) and ref. 25 (o and dashed line).

© 2002 Nature Publishing Group http://neurosci.nature.com

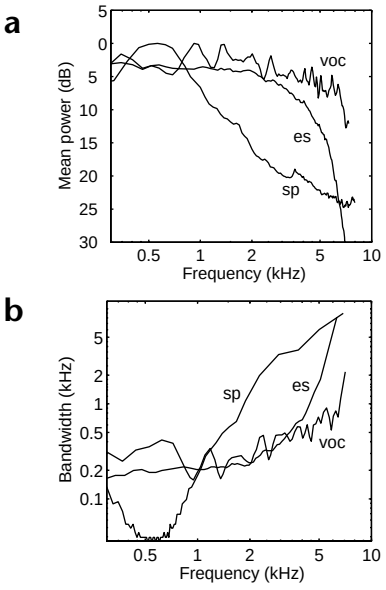


Fig. 8. Predicted bandwidth versus frequency curves assuming equalization of spectral power across bandwidths. (**a**) Average power spectra for environmental sounds (es), vocalizations (voc) and speech (sp). (**b**) The corresponding bandwidth curves are chosen so that each spectral band has equal average power. The curves are essentially a reflection of the power spectra. The change in bandwidth over frequency does not show the power law relationships seen in Fig. 7a.

bands, presumably to compensate for the limited precision of each channel. The curves are different from those of efficient coding (**Fig. 7**), because the predictions derived from the average power spectrum ignore the temporal structure of the signal.

Discussion

The main insight provided by this analysis is not into the response properties of individual cochlear fibers, but into how sound is encoded by the specific distribution of response properties of the population. That the filters in an efficient code of sound are localized in time–frequency space is not surprising for general classes of sounds. If particular harmonic structures do not dominate the statistics, then one would expect localized filters. Given localized filters, the only remaining degrees of freedom (beyond the details of the filter form) lie in the particular trade-off between time and frequency in the population. A common mischaracterization of the peripheral auditory system is that it performs a Fourier analysis of the auditory signal. If this were true, filter bandwidth would remain roughly constant as a function of center frequency, which is not observed experimentally. Another approximation of auditory nerve filters is that they have constant sharpness, as in a wavelet representation. This too is inconsistent with the experimental data, which shows a sublinear power law trend of filter sharpness versus center frequency. The results provide an explanation for the distribution of cochlear tuning properties in terms of efficient coding, and suggest that the auditory system is adapted for the efficient representation of a broad range of natural sounds, including sounds in the natural environment and animal vocalizations.

Linear models of auditory coding based on revcor filters can account only for cochlear nerve fiber responses to stimuli within a limited dynamic range, and do not capture effects such as adaptation or two-tone inhibition[4]. The limitations of our simple, but mathematically tractable, model should also be noted. First, this model is linear, so it cannot account for dynamic aspects of auditory filters, such as any form of gain control or broadening of bandwidths for higher stimulus intensities. Second, statistical regularities on a larger time scale (for example, phonemes) cannot be captured because the model optimizes only the efficiency of the code within the analysis window. When the autocorrelations of the different sound ensembles were analyzed, most of the significant temporal correlation fell within the length of the analysis window (unpublished data). Incorporating these factors into a model of auditory periphery would likely increase the efficiency of the sensory code[29–31].

Earlier studies of natural sounds[26,27] have reported that sound classes such as speech and music often obey a statistical regularity of having approximately a $1/f$ power spectrum, which is due to correlations that exist over many time scales. However, these observations were reported for much larger time scales (0.001—10 Hz) than those analyzed here, complementing rather than contradicting our results. Here we provide a characterization of the statistical regularities on a time scale (~0.5—8 kHz) more relevant to the time scales coded at the level of the auditory nerve.

The predicted filter shapes typically do not show the temporal asymmetry of the auditory nerve filters, which have a rapid rise followed by a slower decay. This is because, for simplicity, the model was designed to encode only the signal within the analysis window and filters were not restricted to be causal. Deriving efficient codes for such models is beyond the scope of current methods of ICA. A causal model in which the encoding proceeds forward in time would lead to an asymmetry in the filters because structure preceding the current time would already have been accounted for earlier in the coding. The filters in such a code could presumably capture the structure of the onset or offset envelopes of natural sounds more closely.

The result that the optimal representation changes depending on the class of sounds suggests that similar adaptations might be used by biological systems depending on ecological niche. An auditory system adapted to process a broad class of sounds fits the auditory sensory code described here. If we allow that the human cochlear nerves follow filter bandwidth and sharpness curves similar to those measured physiologically, this analysis also suggests that the acoustic properties of speech make efficient use of the bandwidth available in the auditory system. Efficient representation of speech is nearly identical to that of natural sounds combined with vocalizations, suggesting an evolutionary adaptation of speech to make maximally efficient use of the coding properties of a prelinguistic auditory system.

These results nicely mirror the results on efficient coding of natural images by demonstrating that efficient coding of natural sounds can explain many of the sensory coding properties of the auditory system. Efficient coding of natural scenes results in a population of localized, oriented Gabor wavelet-like filters[11,12]. The auditory equivalent is the gammatone filter, or a gamma-modulated sinusoid. A prevalent form of structure in natural scenes is an edge that can be efficiently encoded by a population of Gabor filters. Similarly, sound onsets or 'acoustic edges' can be efficiently encoded by a population of filters that resembles a gammatone filter bank. In both cases, however, the interpretation offered by the theory is not that these filters are 'edge detectors', but rather that the code is optimized for a more general class of patterns: those with edges and those that vary smoothly. These results lend further support to the hypothesis that efficient coding is a general

© 2002 Nature Publishing Group http://neurosci.nature.com

© 2002 Nature Publishing Group http://neurosci.nature.com

principle for sensory coding. A challenge that remains is finding workable models that can account for the many types of non-linearities and higher-order processing in perceptual systems.

METHODS

Theory. The relationship between efficient coding and statistical density estimation can be seen by applying Shannon's source coding theorem, which states that the lower bound on the expected code length is the entropy of the data, $H(p) = -\Sigma p(x) \log p(x)$. The true probability density $p(x)$, however, is unknown and must be approximated by the density $q(x)$ assumed by the model. From this, the lower bound on expected code length $l(x)$ becomes

$$E[l(x)] \geq \sum_x p(x) \log \frac{1}{q(x)} = \sum_x p(x) \log \frac{1}{p(x)} + \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (2)$$

The term on the left is the entropy; the term on the right is the Kullback-Leibler divergence between p and q and is zero if and only if $p = q$. Thus, the closer q is to p, the lower the bound on the expected code length.

To obtain an expression for the likelihood of the data under the model (1), the stimulus waveform $\mathbf{x}$ is expressed as a sum of basis functions ϕ_i weighted by coefficients a_i, $\mathbf{x} = \Sigma_i a_i \phi_i$, where $\mathbf{x} = [x(1), \ldots, x(N)]$. To obtain the filters h_i, the previous equation is written in matrix form $\mathbf{x} = [\phi_i, \ldots, \phi_k]\mathbf{a} = \Phi\mathbf{a}$. Then $\mathbf{a} = \Phi^{-1}\mathbf{x}$, yielding the (FIR) filters in the rows of Φ^{-1}. The data likelihood is then $p(\mathbf{x} \mid \Phi) = p(\mathbf{a}) / \mid \det \Phi \mid$ (ref. 32,33).

Data. The ensemble of environmental sounds were obtained from a variety of sources, including rustling brush, crunching leaves and twigs, rain, fire, and forest and stream sounds, and was 45 s in total duration. Animal vocalizations were selected from a collection of mammalian vocalizations[34]. A representative sample of 44 vocalizations was used, and periods of silence were removed. Although the collection was recorded to isolate the vocalizations, the background sounds of the natural habitat, typically birds and insects, were audible in many recordings. Speech was obtained from the TIMIT continuous speech corpus using 100 male and female speakers. To reduce the amount of computation required, the environmental and animal vocalization stereo sounds were converted to mono and down-sampled from 44.1 kHz to 14.7 kHz. Speech was used at the original sampling frequency of 16 kHz. All waveforms were highpass filtered using a cutoff equal to the sampling frequency divided by the window size, which was 115 Hz for environmental sounds and animal vocalizations and 125 Hz for speech. Datasets were then constructed with 128-sample segments, randomly selected from the sound ensembles.

Algorithm. The details of the basic algorithm to derive efficient codes have been given previously[12,14]. Briefly, the basis matrix Φ (or, equivalently, the filter set) is optimized by maximizing the likelihood of the data ensemble under the model

$$\Delta\Phi \propto \Phi\Phi^T \frac{\partial}{\partial\Phi} \log p(\mathbf{x} \mid \Phi) = \Phi(\mathrm{I} - \varphi(\mathbf{a})\mathbf{a}^T) \quad (3)$$

where the prefactor $\Phi\Phi^T$ is used for faster convergence[35], and $\varphi(\mathbf{a}) = (\log p(\mathbf{a}))'$. Independence of the coefficients is assumed, as in $p(\mathbf{a}) = \Pi_i p(a_i)$. The distribution $p(a_i)$ is typically fixed *a priori*, but here we fit this distribution to the data using a generalized Gaussian[36–38], $p(a) \propto \exp(-\mid a \mid^q)$. This yields $\varphi(a) = -\theta \mid a - \mu \mid^{q-1} qc\sigma^{-q}$, where $\theta = \mathrm{sign}(s)$ and $c = [\Gamma(3/q)/\Gamma(1/q)]^{q/2}$ (subscripts omitted for clarity). By inferring the maximum *a posteriori* value of q_i from the data, the model can fit a broad range of statistical distributions, including those assumed by both principal and independent component analysis, and was very well matched to the distributions observed here. It should be noted that the learning rule is not intended to be biologically plausible but simply a method for deriving the theoretical predictions of efficient coding for the ensembles of natural sounds.

In all experiments presented here, five optimizations were performed from different random initial conditions of Φ, all yielding qualitatively similar results. During optimization each gradient step was estimated using a block of 640 waveform segments. The gradient step size was reduced linearly from 10^{-1} to 10^{-5} over 10,000 iterations and then optimized further at the final step size for an additional 10,000 iterations. Each generalized Gaussian parameter q_i for the output distribution $p(a_i)$ was estimated throughout optimization, after every 20,000 patterns. Instability in the gradient can result from small values q_i, so the contribution of the q_i to the gradient was limited to a minimum of two-thirds during optimization.

Time–frequency analysis. To determine the extent of a filter in time–frequency space, the temporal extent was measured using the width required to cover 95% of the filter power. Frequency width was measured using the spectral bandwidth at 10 dB down from the peak. Filters that were not localized (where the main spectral peak at accounted for less than 50% of the total power) were omitted from the plot and totaled 7, 1 and 0 out of 128 filters for environmental sounds, vocalizations and speech, respectively.

Acknowledgements

The author thanks C. Olson, B. Olshausen and L. Holt for discussions and feedback on the manuscript.

Competing interests statement

The author declares that he has no competing financial interests.

RECEIVED 15 OCTOBER 2001; ACCEPTED 8 FEBRUARY 2002

1. Barlow, H. B. Possible principles underlying the transformation of sensory messages. in *Sensory Communication* (ed. Rosenbluth, W. A.) 217–234 (MIT Press, Cambridge, 1961).
2. Kiang, N. Y.-S., Watanabe, T., Thomas, E. C. & Clark, L. F. *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (MIT Press, Cambridge, Massachusetts, 1965).
3. Evans, E. F. Frequency selectivity at high signal levels of single units in cochlear nerve and nucleus. in *Psychophysics and Physiology of Hearing* (eds. Evans, E. F. & Wilson, J. P.) 185–192 (Academic, New York, 1977).
4. de Boer, E. & de Jongh, H. R. On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *J. Acoust. Soc. Am.* **63**, 115–135 (1978).
5. Carney, L. H. & Yin, T. C. T. Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *J. Neurophys.* **60**, 1653–1677 (1988).
6. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Optical Soc. Am. A* **12**, 2379–2394 (1987).
7. Field, D. J. What is the goal of sensory coding? *Neural Comp.* **6**, 559–601 (1994).
8. Linsker, R. Perceptual neural organization—some approaches based on network models and information-theory. *Annu. Rev. Neuro.* **13**, 257–281 (1990).
9. Atick, J. J. Could information-theory provide an ecological theory of sensory processing. *Network Comp. Neural Sys.* **3**, 213–251 (1992).
10. Rieke, F., Bodnar, D. A. & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. B Biol. Sci.* **262**, 259–265 (1995).
11. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
12. Bell, A. J. & Sejnowski, T. J. The 'independent components' of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338 (1997).
13. van Hateren, J. H. & Ruderman, D. L. Independent component analysis of natural images sequences yield spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond. B Biol. Sci.* **265**, 2315–2320 (1998).
14. Lewicki, M. S. & Olshausen, B. A. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A* **16**, 1587–1601 (1999).
15. Comon, P. Independent component analysis, a new concept? *Signal Process.* **36**, 287–314 (1994).
16. Bell, A. J. & Sejnowski, T. J. An information maximization approach to blind separation and blind deconvolution. *Neural Comp.* **7**, 1129–1159 (1995).

17. Laughlin, S. B. Matching coding to scenes to enhance coding efficiency. in *Physical and Biological Processing of Images* (eds. Braddick, O. J. & Sleigh, A. C.) 42–72 (Springer, Berlin, 1983).
18. Bell, A. J. & Sejnowski, T. J. Learning the higher-order structure of a natural sound. *Netw. Comput. Neural Syst.* **7**, 261–267 (1996).
19. Irino, T. & Patterson, R. D. A time-domain, level-dependent auditory filter: the gammachirp. *J. Acoust. Soc. Am.* **101**, 412–419 (1997).
20. Zoharian, A. S. & Rothenberg, M. Principal-component analysis for low redundancy encoding of speech spectra. *J. Acoust. Soc. Am.* **69**, 832–845 (1981).
21. Mallat, S. *A Wavelet Tour of Signal Processing* 2nd edn. (Academic, London, 1999).
22. Lewicki, M. S. & Sejnowski, T. J. Learning overcomplete representations. *Neural Comput.* **12**, 337–365 (2000).
23. Moore, B. C. J. (ed.) *Frequency Selectivity in Hearing* (Academic, London, 1986).
24. Evans, E. F. Cochlear nerve and cochlear nucleus. in *Handbook of Sensory Physiology* Vol. 5/2 (eds. Keidel, W. D. & Neff, W. D.) 1–108 (Springer, Berlin, 1975).
25. Rhode, W. S. & Smith, P. H. Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hearing Res.* **18**, 159–168 (1985).
26. Voss, R. F. & Clarke, J. 1/f noise in music and speech. *Nature* **258**, 317–318 (1975).
27. Attias, H. & Schreiner, C. Low-order temporal statistics of natural sounds. in *Advances in Neural and Information Processing Systems* Vol. 9 (Morgan Kaufmann, San Mateo, California, 1997).
28. Furth, P. M. & Andreou, A. G. A design framework for low power analog filter banks. *IEEE Trans. Circuits Syst. I* **42**, 966–971 (1995).
29. Lewicki, M. S. & Sejnowski, T. J. Coding time-varying signals using sparse, shift-invariant representations. in *Advances in Neural Information Processing Systems* Vol. 11, 730–736 (MIT Press, Cambridge, Massachusetts, 1999).
30. Brenner, N., Bialek, W. & van Steveninck, R. D. Adaptive rescaling maximizes information transmission. *Neuron* **26**, 695–702 (2000).
31. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nat. Neurosci.* 4, 819–825 (2001).
32. Pearlmutter, B. A. & Parra, L. C. A context-senstive generalization of ICA. in *Proceedings of the International Conference on Neural Information Processing* 151–157 (Springer, Singapore, 1996).
33. Cardoso, J.-F. Infomax and maximum likelihood for blind source separation. *IEEE Signal Process. Lett.* **4**, 109–111 (1997).
34. Emmons, L. H., Whitney, B. M. & Ross, D. L. Sounds of the neotropical rainforest mammals [audio CD] (Library of Natural Sounds, Cornell Laboratory of Ornithology, Ithaca, New York, 1997).
35. Amari, S., Cichocki, A. & Yang, H. H. A new learning algorithm for blind signal separation. in *Advances in Neural and Information Processing Systems* Vol. 8, 757–763 (Morgan Kaufmann, San Mateo, California, 1996).
36. Box, G. E. P. & Tiao, G. C. *Bayesian Inference in Statistical Analysis* (Addison-Wesley, Reading, Massachusetts, 1973).
37. Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693 (1989).
38. Simoncelli, E. P. & Adelson, E. H. Noise removal via Bayesian wavelet coring in *Proceedings of the 3rd IEEE International Conference on Image Processing* Vol. 1, 379–382 (IEEE Signal Processing Society, Lausanne, 1996).

© 2002 Nature Publishing Group http://neurosci.nature.com